Multi-CAST

Northern Kurdish
corpus counts

◆

Geoffrey Haig
University of Bamberg

Hanna Thiele
University of Bamberg / University of Freiburg

May 2019
v2.0

ARC CENTRE OF EXCELLENCE FOR
THE DYNAMICS OF LANGUAGE

Australian Government
Australian Research Council

University of Bamberg

DFG

Multi-CAST

Multilingual Corpus of
Annotated Spoken Texts

Contents

1   Notes on the GRAID counts

This document collects tables with frequency counts for combinations of selected GRAID symbols in version 1905 (from May 2019) of the Multi-CAST Northern Kurdish corpus. The tables are intended to offer cursory impressions of the relative proportions between different types of referring expression; they do not provide exact summaries of the annotations.

Only a small number of basic GRAID symbols are counted:

Function symbols

⟨0⟩	zero
⟨pro⟩	definite pronoun
⟨np⟩	full noun phrase
⟨other⟩	form not further specified

Person/Animacy symbols

⟨.1⟩	first person
⟨.2⟩	second person
⟨.h⟩	third person, human
⟨.d⟩	third person, anthropomorphic
ø	third person, non-human

Function symbols

⟨:a⟩	subject of a transitive clause
⟨:s⟩	subject of an intransitive clause
⟨:ncs⟩	non-canonical subject
⟨:p⟩	direct object
⟨:obl⟩	oblique argument
⟨:g⟩	goal argument
⟨:l⟩	locational argument
⟨:poss⟩	possessive
⟨:pred⟩	predicate
⟨:other⟩	function not further specified

Clause boundary symbols

⟨##⟩	independent clause
⟨#⟩	other clause

Only basic categories are listed; categories represented by complex symbols with additional specifiers (e.g. ⟨dem_pro⟩ 'demonstrative pronoun') have been subsumed under the more basic category (e.g. ⟨pro⟩ 'definite pronoun'). Please refer to the annotation notes for this corpus for information on all annotated categories, including those not listed here.

The tables in this document can be recreated with the `mc_table` function from *multicastR*, the companion package to Multi-CAST for the statistical computing language R. Please refer to the package documentation (`?multicastR`, `?mc_table`) for more information.

2   The Northern Kurdish corpus

GRAID		⟨:a⟩	⟨:s⟩	⟨:ncs⟩	⟨:p⟩	⟨:obl⟩	⟨:g⟩	⟨:l⟩	⟨:poss⟩	⟨:pred⟩	⟨:other⟩	*totals*
⟨0	.1⟩	19	8	0	11	0	1	0	0	0	0	39
⟨0	.2⟩	18	13	0	0	0	0	0	0	0	0	31
⟨0	.h⟩	200	233	0	44	0	0	0	1	0	0	478
⟨0	.d⟩	0	2	0	0	0	0	0	0	0	0	2
⟨0⟩		7	44	0	82	0	2	0	0	0	0	135
⟨pro	.1⟩	70	67	0	11	10	11	0	58	0	0	227
⟨pro	.2⟩	30	41	0	5	9	10	0	19	0	0	114
⟨pro	.h⟩	22	41	0	5	15	18	0	55	0	0	156
⟨pro	.d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨pro⟩		0	3	0	2	5	6	5	13	0	0	34
⟨np	.h⟩	39	86	0	22	28	20	1	76	32	25	329
⟨np	.d⟩	0	1	0	0	0	0	0	0	0	0	1
⟨np⟩		6	122	0	211	33	114	59	37	30	45	657
⟨other	.h⟩	0	0	0	1	1	0	0	0	0	0	2
⟨other	.d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨other⟩		1	1	0	14	0	27	11	2	87	0	143
totals		412	662	0	408	101	209	76	261	149	70	
⟨##⟩												269
⟨#⟩												831
totals												1100

Table 1   Summarized GRAID counts for the entire Northern Kurdish corpus.

2.1 muserz01

GRAID		⟨:a⟩	⟨:s⟩	⟨:ncs⟩	⟨:p⟩	⟨:obl⟩	⟨:g⟩	⟨:l⟩	⟨:poss⟩	⟨:pred⟩	⟨:other⟩	*totals*
⟨0	.1⟩	12	7	0	11	0	0	0	0	0	0	30
⟨0	.2⟩	11	10	0	0	0	0	0	0	0	0	21
⟨0	.h⟩	98	123	0	24	0	0	0	1	0	0	246
⟨0	.d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨0⟩		1	11	0	27	0	0	0	0	0	0	39
⟨pro	.1⟩	44	49	0	8	6	8	0	40	0	0	155
⟨pro	.2⟩	20	27	0	5	7	4	0	12	0	0	75
⟨pro	.h⟩	13	22	0	2	14	14	0	30	0	0	95
⟨pro	.d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨pro⟩		0	0	0	0	2	5	4	7	0	0	18
⟨np	.h⟩	22	43	0	13	20	10	0	44	30	21	203
⟨np	.d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨np⟩		3	77	0	138	22	56	42	13	20	23	394
⟨other	.h⟩	0	0	0	1	1	0	0	0	0	0	2
⟨other	.d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨other⟩		0	0	0	3	0	17	10	0	65	0	95
totals		224	369	0	232	72	114	56	147	115	44	
⟨##⟩												148
⟨#⟩												470
totals												618

Table 2 Summarized GRAID counts for the *muserz01* text.

2.2 muserz02

GRAID		⟨:a⟩	⟨:s⟩	⟨:ncs⟩	⟨:p⟩	⟨:obl⟩	⟨:g⟩	⟨:l⟩	⟨:poss⟩	⟨:pred⟩	⟨:other⟩	totals
⟨0	.1⟩	7	1	0	0	0	1	0	0	0	0	9
⟨0	.2⟩	7	3	0	0	0	0	0	0	0	0	10
⟨0	.h⟩	102	110	0	20	0	0	0	0	0	0	232
⟨0	.d⟩	0	2	0	0	0	0	0	0	0	0	2
⟨0⟩		6	33	0	55	0	2	0	0	0	0	96
⟨pro	.1⟩	26	18	0	3	4	3	0	18	0	0	72
⟨pro	.2⟩	10	14	0	0	2	6	0	7	0	0	39
⟨pro	.h⟩	9	19	0	3	1	4	0	25	0	0	61
⟨pro	.d⟩	0	0	0	0	0	0	0	0	0	0	0
⟨pro⟩		0	3	0	2	3	1	1	6	0	0	16
⟨np	.h⟩	17	43	0	9	8	10	1	32	2	4	126
⟨np	.d⟩	0	1	0	0	0	0	0	0	0	0	1
⟨np⟩		3	45	0	73	11	58	17	24	10	22	263
⟨other .h⟩		0	0	0	0	0	0	0	0	0	0	0
⟨other .d⟩		0	0	0	0	0	0	0	0	0	0	0
⟨other⟩		1	1	0	11	0	10	1	2	22	0	48
totals		188	293	0	176	29	95	20	114	34	26	
⟨##⟩												121
⟨#⟩												361
totals												482

Table 3 Summarized GRAID counts for the *muserz02* text.

Multi-CAST

Multilingual Corpus of Annotated Spoken Texts

multicast.aspra.uni-bamberg.de/